J. Math. and Its Appl.
ISSN: 1829-605X
Vol. 6, No. 1, May 2009, 63–76

PENGKLASTERAN DATA KATEGORIS DENGAN ALGORITMA *SHARED NEAREST NEIGHBOR*

Alvida Mustikarukmi[1], M. Isa Irawan[2], Nurul Hidayat[3]
Jurusan Matematika, Institut Teknologi Sepuluh Nopember
[1]alvida_yusuf@yahoo.co.id

Abstrak

Pengklasteran objek data merupakan salah satu cara untuk mempermudah dalam membaca data, terutama data berdimensi tinggi. Obyek-obyek data berada dalam satu klaster jika mempunyai kesamaan yang tinggi, dan sebaliknya, berada pada klaster berbeda jika menunjukkan ketidaksamaan. Data kategoris merupakan jenis data yang sering digunakan pada database/dataset. Data teks merupakan salah satu data kategoris. Pengklasteran dengan algoritma *shared nearest neighbor* (SNN) didasarkan pada anggapan bahwa titik-titik akan berada dalam klaster yang sama jika jumlah *shared nearest neighbor* melebihi ambang batas yang ditentukan. Algoritma SNN mampu memberikan hasil pengklasteran data teks dengan baik, dimana teks dengan tingkat kesamaan yang ditentukan, akan berada pada klaster yang sama.
Katakunci: *shared nearest neighbor, pengklasteran, data teks.*

1. Pendahuluan

Pengklasteran objek data merupakan salah satu cara untuk mempermudah dalam membaca data, terutama data berukuran besar. Analisis klaster [3]

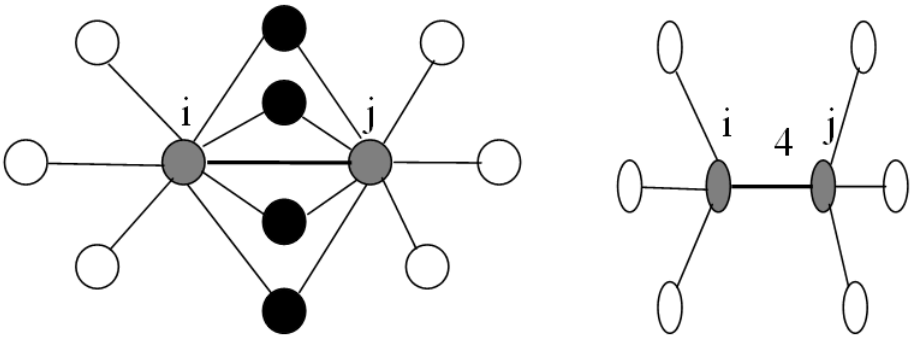
membagi data menjadi grup-grup klaster. Obyek-obyek data berada dalam satu klaster jika mempunyai kesamaan yang tinggi, dan sebaliknya, berada pada klaster berbeda jika menunjukkan ketidaksamaan. Ukuran kesamaan dapat berdasarkan jarak antar obyek data jika dipandang dalam perspektif geometri. Semakin dekat jarak antar obyek data, semakin menunjukkan kesamaan. Klaster-klaster yang terbentuk dapat menunjukkan ukuran, bentuk, dan kepadatan yang berlainan. Metode pengklasteran juga harus disesuaikan dengan jenis data yang akan diklaster. Pengklasteran teks merupakan salah satu teknik klasifikasi takterbimbing *(unsupervised classification)* kemiripan dokumen ke dalam grup-grup yang berbeda. Pengklasteran dokumen adalah cara untuk mengenali klaster atau grup dokumen yang menggunakan bersama fitur. Algoritma pengklasteran teks terdiri dari tiga aspek : penyajian dokumen, ukuran kedekatan dokumen, dan pengurangan dimensi tinggi. Kedekatan dua dokumen diukur berdasarkan pemakaian bersama deretan frekuensi (arti) kata. Pengklasteran koleksi buku menggunakan teknik pengklasteran dokumen karena pengambilan teks pada judul buku diperlakukan sebagai istilah(kata) untuk acuan pengklasteran. Pengklasteran koleksi buku berdasarkan judul buku mencerminkan muatan isi buku.

2. Algoritma *Shared Nearest Neighbor*

Algoritma *shared nearest neighbor* (SNN) terhadap kesamaan pada proses pengklasteran dikenalkan sebagai cara untuk mengatasi masalah pengukuran jarak dalam data berdimensi tinggi [4] dan dikembangkan oleh [5]. Algoritma SNN menemukan *nearest neighbor* masing-masing titik data dan algoritma Jarvis-Patrick mendefinisikan kembali kesamaan antara pasangan titik yang merupakan *shared nearest neighbor* dari dua titik dengan menemukan *k-nearest neighbor* semua titik. Pasangan titik diletakkan dalam klaster yang sama jika :

1. dua titik sebarang menggunakan bersama lebih dari ε *nearest neighbor*

2. dua titik tersebut saling berada dalam daftar *k-nearest neighbor*



Gambar 1: SNN(i, j) sebesar 4

Parameter yang digunakan algoritma SNN dapat mengendalikan resolusi pengklasteran dan memudahkan user untuk mengendalikan berapa banyak titik yang diklaster atau kebalikannya yakni berapa banyak titik yang dikategorikan sebagai noise, adalah sebagai berikut :

1. k daftar neighbor
2. ε, berupa jari-jari(radius) daerah neighbor (Nε(p))dari sebuah titik data p
3. MinT, jumlah minimal titik interior dalam daerah Nε(p)

Langkah–langkah pengerjaan algoritma SNN dinyatakan berikut ini :

1. Bangun matriks kesamaan. Elemen matriks ini berkorespondensi dengan graf kesamaan dimana node berupa titik data dan bobot edge berupa kesamaan antara titik data.
2. Melakukan *sparsing* matriks kesamaan dengan cara mempertahankan k neighbor yang paling mirip saja. Hal ini berkorespondensi dengan cara mempertahankan k *link* terkuat dari graf kesamaan.
3. Bangun graf *shared nearest neighbor* dari matriks kesamaan yang telah dilakukan *sparsing* tersebut.
4. Temukan kepadatan SNN di setiap titik dan titik utama. Kepadatan SNN sebuah titik adalah jumlah titik yang mempunyai kesamaan

SNN lebih besar dari ε, sedangkan kepadatan titik utama adalah semua titik yang mempunyai kepadatan SNN lebih besar dari MinT.

5. Bentuklah klaster-klaster dari titik utama. Jika dua titik utama berada pada dalam radius ε, maka keduanya ditempatkan dalam cluster yang sama

3. Proses Pembentukan Klaster

3.1. Pembobotan Kata

Frekuensi Kata (Term Frequency/TF) menyatakan banyaknya kemunculan sebuah kata dalam sebuah dokumen [1]. Kata yang mempunyai TF tinggi pada sebuah dokumen lebih mengarah terhadap topik dokumen. Pada pemberian bobot kata dilakukan normalisasi sepanjang keseluruhan koleksi dokumen, yaitu :

$$tf_{ij} = \frac{f_{ij}}{\max\{f_{ij}\}} \tag{1}$$

dengan :
f_ij : frekuensi kata i dalam dokumen j

Frekuensi Dokumen (Document Frequency/DF) menyatakan kemunculan sebuah kata pada beberapa dokumen yang berlainan. Invers Frekuensi Dokumen (Inverse Document Frequency/IDF) adalah kebalikan dari DF, yakni kurang mengindikasikan sebuah topik. Bobot IDF digunakan untuk menyajikan spesifisitas kata, jadi bobot tinggi berarti sebuah kata bersifat spesifik pada sebuah dokumen, sedang bobot rendah berarti sebuah kata bersifat umum pada beberapa dokumen. Nilai IDF, invers frekuensi dokumen terhadap kata i, w_i dinyatakan

$$idf_i = {}^2\log(\frac{N}{df_i}) \tag{2}$$

dengan :
d_fi : frekuensi dokumen untuk kata i (banyaknya dokumen yang memuat kata i)
N : jumlah total dokumen

Pembobotan TF/IDF didapatkan dari persamaan

$$w_{ij} = tf_{ij} \cdot idf_i \cdot {}^2\log(\frac{N}{df_i}) \tag{3}$$

dengan : w_i : bobot pada kata i

3.2. Pembentukan Matriks Kemiripan

Entry pada matrik kemiripan diperoleh dari ukuran kemiripan yang digunakan. Ukuran kemiripan data kategoris berdasarkan jarak menggunakan koefisien Jaccard atau ukuran kosinus. Jarak Jaccard dinyatakan pada persamaan (4)

$$d_{ij} = \frac{p}{q + r - p} \tag{4}$$

dengan :
p : banyaknya peubah positif pada kedua obyek data
q : banyaknya peubah positif pada obyek ke-i dan negatif pada obyek ke-j
r : banyaknya peubah negatif pada obyek ke-i dan positif pada obyek ke-j
d : jarak kedua obyek
Sedangkan ukuran kosinus dapat diperoleh dari persamaan berikut :

$$sama(d_i, d_j) = \frac{d_i \cdot d_j}{|d_i||d_j|} = \frac{\sum_{k-1}^{n} c_{i,k} c_{j,k}}{\sqrt{\sum_{k-1}^{n} c_{i,k}^2} \sqrt{\sum_{k-1}^{n} c_{j,k}^2}} \tag{5}$$

dengan :
d_i : vektor data ke-i
d_j : vektor data ke-j
n : banyaknya data
$c_{i,k}$: skalar pada vektor data ke-i
$c_{j,k}$: skalar pada vektor data ke-j

3.3. Pemrosesan Data

Sumber data mengambil dataset pada perpustakaan pusat ITS berupa koleksi buku berbahasa Indonesia sebanyak 500 buah sebagai obyek yang akan dikaster. Pengklasteran koleksi buku dilakukan sesuai dengan topik muatan buku, dimana kata-kata pada judul buku diambil sebagai kata kunci. Untuk pengindeksan dokumen dilakukan langkah-langkah preprocessing berikut :

1. Dilakukan proses memecah teks penyusun judul buku dari masing-masing buku menjadi kata, frasa, simbol, atau unsur-unsur bermakna lain yang disebut token.

2. Penghilangan kata depan, sandang kata sambung, dan jenis kata yang tidak termasuk kata significant (removing stop-word); tanda baca (punctuation); spasi

Artikel ini mengabaikan relasi terhadap kata sinonim dan polisemi.

4. Desain Sistem

Algoritma SNN membutuhkan parameter k untuk menyatakan jumlah obyek *nearest neighbor* [2]. Pada pengklasteran ini kesamaan antar buku ditentukan oleh kesamaan kata- kata pada judul buku. Ukuran kesamaan kata berdasarkan frekuensi kemunculan kata-kata pada buku yang sama. Pembentukan kata kunci untuk membuat kategori berdasarkan nilai df_i setiap kata. Kata dengan nilai df_i maksimal digunakan sebagai acuan pembentukan kata kunci. Pengambilan $l$kata sebagai kata kunci dimaksudkan untuk kejadian berulang, dimana l kata muncul bersamaan dalam sebuah judul buku. Kesamaan kata penyusun judul buku yang muncul di antara dua buku menyiratkan kesamaan topik antar kedua buku. Banyaknya l kata yang digunakan bersama oleh dua buku, menyiratkan nilai SNN antar dua obyek data dengan ketentuan :

1. Jumlah kata yang sama disimpan sebagai jumlah SNN, jika tidak ada NN yang sama, maka SNN menunjukkan nol

2. Jika jumlah SNN antar obyek (buku) = l ,simpan (w_i, w_j) sebagai kata kunci

Pasangan kata yang muncul berulang (*co-occurence*) pada sebuah teks menandakan kata kunci [7]. Judul dari beberapa buku yang memuat kata kunci yang sama menunjukkan kedekatan sehingga pasangan kata tersebut dapat dijadikan acuan pembentukan kategori. Pasangan(T_i, T_j) menyatakan pasangan kata yang muncul pada judul buku. Frekuensi munculnya pasangan ini menandai kedekatan antar buku, sehingga kedua kata berada dalam satu kategori. Dengan cara yang sama, bandingkan kata T_i dengan

lainnya, kumpulkan dalam satu kategori. Buku-buku diletakkan dalam satu kategori jika memuat kata kunci yang sama, selanjutnya, membentuk matriks *Buku-kata* untuk setiap kategori. Elemen matrik *Buku-kata* diisi bobot yang diperoleh dari persamaan Jaccard.

Nilai parameter MinT diambil dari jumlah minimal buku dalam sebuah kategori. Buku-buku tersebut akan membentuk sebuah klaster. Beberapa buku yang tidak masuk kategori dimasukkan sebagai *noise*, jika kata – kata penyusun buku tidak mempunyai keterkaitan dengan buku lain. Hal ini ditandai dengan nilai dfi kata kuncinya rendah, misal nilai $df_i = 1$ atau $df_i = 2$.

Matrik buku-kata direpresentasikan sebagai matrik A kemudian dilakukan langkah-langkah berikut ini :

$A \quad \rightarrow A^T$
$AA^T \rightarrow B$
$B \quad \rightarrow C$, dimana matrik B menyatakan matrik kesamaan buku-buku.

Sifat Matriks Kesamaan adalah:

1. tak negatif : $m_{ij} \geq 0$
2. refleksif : $m_{ij} = 1$ untuk semua $i = j$
3. simetri : $m_{ij} = m_{ji}$
4. terbatas : $0 \leq m_{ij} \leq 1$

Kemudian setiap elemen matrik B diberi bobot wii dengan menggunakan persamaan Jaccard. jadi elemen-elemen matrik B memuat angka 1 dan 0. Matrik C diperoleh dari transformasi matrik B dengan menggunakan koefisien Jaccard $c_{ij} = b_{ij}/(b_{ii} + b_{jj} - b_{ij})$. Koefisien tersebut mengestimasi derajat kesamaan antara pasangan *neighbor*. Karena pasangan *neighbor* terdekat digunakan untuk menghasilkan klaster, maka matrik C menyatakan matrik kesamaan pasangan dua buku (b_i, b_j). Baris i pada matrik C dinyatakan sebagai ketetanggaan kata i. Matrik C menunjukkan bahwa dua buku (b_i, b_j) mempunyai kesamaan yang paling dekat jika bobot (b_i, b_j) menunjukkan nilai tertinggi.

4.1. Penghitungan Frekuensi dan Pemilahan Token

Pada pengklasteran ini mengambil nilai $k = 6$ yang menyatakan banyaknya kata yang digunakan pada kejadian ulang kata *(co-occurence word)*. Nilai k menyatakan jumlah kolom hasil tokenisasi yang disimpan pada tabel Books. Jumlah kolom tabel Books terdiri dari 8 kolom. Kolom pertama berisi ID_Buku, kolom kedua memuat judul buku, dan 6 kolom berikutnya menyatakan hasil token. Jika judul buku yang disusun kurang dari 6 kata, maka kolom sisa bernilai null, contoh : buku "Aljabar Linier" memuat 2 token sehingga 4 kolom berikut bernilai null. Sedangkan judul buku yang memuat 8 token, hanya dibaca maksimum 6 token.

Tabel 4.1.1. Tabel Books disertai Hasil Tokenisasi

Books							
BookID	**Title**	**T1**	**T2**	**T3**	**T4**	**T5**	**T6**
8001	Adat dan upacara perkawinan daerah Bengkulu	daerah	Adat	perkawinan	upacara	Bengkulu	
8002	Adat dan upacara perkawinan daerah Daerah Jambi	daerah	Adat	perkawinan	upacara	Jambi	
8003	Adat dan upacara perkawinan daerah Istimewa Aceh	daerah	Adat	perkawinan	upacara	Istimewa	Aceh
8004	Adat dan upacara perkawinan daerah Istimewa Yogyakarta	daerah	Adat	perkawinan	upacara	Istimewa	Yogyakarta

4.2. Pembentukan Katakunci dan Kategori

Pada artikel ini parameter SNN mengambil nilai 2, yang menyatakan dua kata digunakan bersama antar buku dan dijadikan sebagai kata kunci untuk membentuk klaster. Banyaknya buku yang memuat dua kata tersebut dihitung dan jumlah buku dijadikan acuan sebagai banyaknya anggota sebuah klaster. Jika parameter MinT menyatakan nilai minimal jumlah buku untuk membentuk sebuah klaster. Nilai MinT = 8 ditetapkan, sebuah ka-

takunci hanya terdapat kurang dari 8 buku, maka tidak dapat membentuk sebuah klaster. Running dengan Matlab pada M-file katakunci.m menampilkan jumlah buku yang memuat kata kunci.

Kata kunci yang dipilih adalah yang memuat sedikitnya 8 buku. Selanjutnya berdasarkan hasil dari file katakunci.m dibangun klaster dengan melakukan running file kategori.m. Penyusunan klaster menandakan topik yang muncul dominan berdasarkan kata kunci yang diguna-kan. Contoh : memebntuk sebuah klaster berdasarkan kata kunci 'manajemen' dan 'konsep'.

```
Buku = Columns 1 through 9
[2] [289] [1x132 char] 'manajemen'   'konsep'      'management'
                       'daya'        'strategis'   'globalisasi'
[2] [286] [1x38  char] 'manajemen'   'konsep'      'strategik'
                       'kasus'           ''             ''
[2] [272] [1x53  char] 'manajemen'   'proyek'      'sampai'
                       'konseptual'  'operational'       ''
[2] [258] [1x47  char] 'manajemen'   'konsep'      'teknik'
                       'modern'      'personalia'       ''
[2] [225] [1x82  char] 'manajemen'   'manajemen'   'bisnis'
                       'kualitas'    'penerapan'   'kualitas'
[2] [175] [1x87  char] 'manajemen'   'konsep'      'konsep'
                       'bisnis'      'penerapan'        ''
[2] [168] [1x26  char] 'manajemen'   'konsep'      'Indonesia'
                            ''           ''             ''
[2] [158] [1x43  char] 'manajemen'   'konsep'      'teori'
                       'berbasis'    'sekolah'          ''
[2] [ 37] [1x52  char] 'manajemen'   'konsep'      'Akuntansi'
                       'rekayasa'    'manfaat'          ''
```

4.3. Pembentukan matrik kesamaan Buku-Buku

Source code pada file Matrikkesamaan.m membangun matrik buku-kata berdasarkan hasil klaster digunakan untuk membangun matrik kesamaan Buku-Buku. Input baris matrik Buku-Kata menyatakan buku, sedang kolom menyatakan kata (token) pada judul buku.

Buku dan kata hasil token dari setiap klaster membentuk matrik buku-kata yang digunakan untuk mengetahui seberapa dekat topik antara dua buku. Setiap elemen (i, j) pada matrik Buku-Kata memuat hasil

pembobotan kata berdasarkan frekuensi kata dengan persamaan (3). Kemudian matrik kesamaan Buku-Buku dibentuk berdasarkan persamaan (4) dan hasilnya berupa, matrik berukuran $(n \times n)$, dimana n adalah jumlah buku dalam sebuah klaster.

Matrik Kesamaan Buku-Buku (9 x 9 sel):

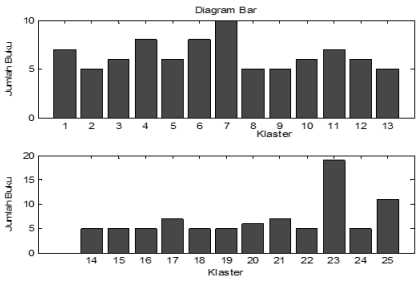
$$\begin{bmatrix} 1 & 0,077 & 0,007 & 0,048 & 0,011 & 0,049 & 0,064 & 0,047 & 0,044 \\ 0,076 & 1 & 0,008 & 0,055 & 0,012 & 0,057 & 0,077 & 0,054 & 0,051 \\ 0,007 & 0,008 & 1 & 0,008 & 0,008 & 0,008 & 0,01 & 0,008 & 0,008 \\ 0,048 & 0,055 & 0,008 & 1 & 0,013 & 0,059 & 0,082 & 0,056 & 0,052 \\ 0,011 & 0,012 & 0,008 & 0,013 & 1 & 0,911 & 0,019 & 0,019 & 0,012 \\ 0,049 & 0,056 & 0,008 & 0,059 & 0,917 & 1 & 0,085 & 0,058 & 0,054 \\ 0,064 & 0,077 & 0,01 & 0,082 & 0,019 & 0,085 & 1 & 0,079 & 0,072 \\ 0,047 & 0,054 & 0,008 & 0,056 & 0,013 & 0,058 & 0,08 & 1 & 0,051 \\ 0,044 & 0,05 & 0,008 & 0,052 & 0,012 & 0,054 & 0,072 & 0,051 & 1 \end{bmatrix}$$

Pembentukan matrik kesamaan buku-buku digunakan untuk menampilkan kedekatan antar buku berdasarkan nilai bobot pada setiap sel matriks. Matriks ini bernilai satu pada semua sel diagonalnya (i, i), karena menyatakan buku itu sendiri. Pada (b_1, b_2) mempunyai bobot kesamaan nilai 0,77 yang menunjukkan nilai kesamaan tertinggi pada baris ke-1. Artinya, bahwa buku ke-1 (berjudul "Manajemen strategi : daya saing dan globalisasi konsep") mempunyai kesamaan tertinggi dengan buku ke-2 (berjudul "Manajemen strategi : konsep dan kasus") dibandingkan dengan 7 buku lain. Kesamaan buku ke-1 terhadap buku ke-2 ditunjukkan dengan penggunakan bersama tiga kata yaitu : manajemen, konsep, dan strategi.

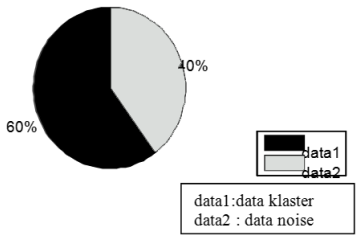
4.4. UJI COBA

Pengujian ini dimaksudkan sebagai alat evaluasi terhadap pemberian nilai parameter selama proses pengklasteran. Jika nilai SNN = 2, kata kunci dibentuk dari dua pasangan kata yang digunakan bersama untuk menyusun judul buku antara 2 buku. Asumsi bahwa kedekatan dua judul buku ditunjukkan oleh penggunaan bersama dua kata antar dua buku. MinT adalah parameter yang digunakan sebagai ambang batas jumlah minimal buku untuk membentuk sebuah klaster.

Berikut ini adalah beberapa hasil dari nilai parameter berbeda untuk membentuk sebuah klaster dari seluruh koleksi buku : i. Nilai SNN= 2, T= 5 Diagram bar pada Gambar 2. menggambarkan hasil pengklas-



Gambar 2: Diagram Bar



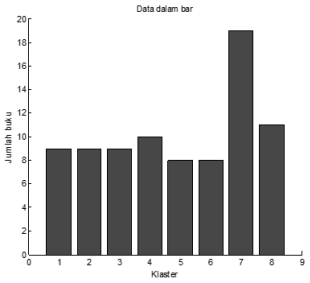
Gambar 3: Data terklaster terhadap Data Noise

tern. Sumbu-x menyatakan identitas klaster (IDKategori), setiap batang bar menunjukkan sebuah klaster dan sumbu-y menyatakan jumlah buku pada setiap klaster. Pada Gambar 3 lingkaran menunjukkan rasio buku yang diklaster terhadap data noise (data tidak terklaster).

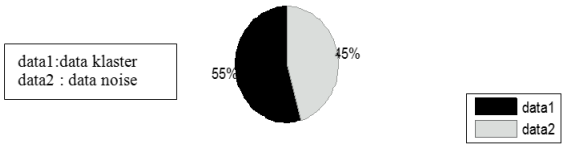
ii. nilai SNN = 2, T = 3 iii. nilai SNN =1, T = 5 iv. nilai SNN = 1, T= 10

4.5. Kesimpulan

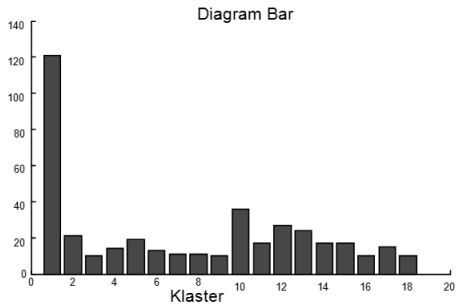
Pengklasteran buku ini didasarkan pada kesamaan kata penyusun judul di antara dua buku yang menyatakan kedekatan topik kedua buku tersebut. Katakunci dipilih berdasarkan nilai frekuensi kemunculan pada koleksi bu-



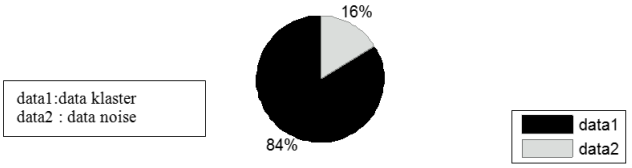
Gambar 4: Diagram Bar



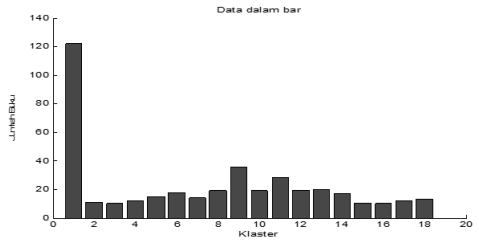
Gambar 5: Data terklaster terhadap Data Noise



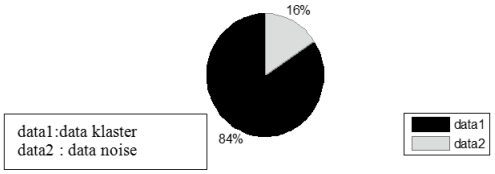
Gambar 6: Diagram Bar



Gambar 7: Data terklaster terhadap Data Noise



Gambar 8: Data terklaster terhadap Data Noise



Gambar 9: Data terklaster terhadap Data Noise

ku dimana frekuensi kata yang tinggi menyiratkan kata bersifat umum, sedangkan frekuensi yang rendah menyiratkan kata bersifat khusus.
Penentuan nilai parameter pada algoritma SNN akan mempengaruhi hasil pengklasteran, yaitu :
Parameter k merupakan input awal dimana penambahan jumlah kata yang akan digunakan pada pengklasteran akan mengakibatkan jumlah buku yang terklaster juga meningkat. Parameter SNN digunakan untuk pembentukan kata kunci. Pembentukan kata kunci dari kata-kata hasil tokenisasi menentukan pembentukan klaster dengan memasukkan buku-buku yang memuat katakunci. Semakin sedikit kata kunci yang dapat digunakan semakin sedikit jumlah klaster. Ada dua kemungkinan :

1. Sebuah klaster memuat buku-buku dalam jumlah besar dan ini akan menyebabkan sebuah klaster dapat memuat buku-buku berbagai topik yang berbeda.

2. Sebuah klaster hanya memuat buku-buku bertopik sama, sehingga jumlah buku yang termasuk ke dalam sebuah klaster hanya sedikit. Hal ini menyebabkan data noise sangat banyak

Untuk parameter T, nilai yang ditetapkan akan mempengaruhi banyaknya

buku dalam sebuah klaster. Penentuan nilai ambang batas yang tepat akan mempangaruhi kualitas klaster. Penggunaan algoritma SNN diharapkan dapat menampilkan hasil pengklasteran dengan kualitas baik untuk mempermudah pembacaan data.

4.6. Saran

Penggunaan model lain yang mendukung seperti penentuan konteks sebagai acuan pengklasteran dokumen dapat dipertimbangkan. Penentuan kata kunci dari kata berfrekuensi tinggi memerlukan kajian lebih lanjut.

Pustaka

[1] Ertoz L., Steinbach M., dan Kumar V.(2002), "Finding topics in document, a *Shared nearest neighbor* Approach Clustering and Information Retrieval", Kluwer Academic Publisher.

[2] Heidelberg (2005), "High Dimensional Shared Narest Neighbor Clustering Algorithm", Lecture Notes on Compuer Science, Publisher Springer Berlin vol.3614, Hal. 494-50

[3] Jain A.K dan Dubes R.C (1988), "Algirthms for Clustering Data", Prentice Hall

[4] Jarvis R.A. dan Patrick E. (1973), "Clustering Using a Similarity Measure based on Shared Nearest Neighbor", *Proceeding IEEE Transaction on Computer*, vol C-22, hal. 1025-1034.

[5] Karphys G., Han E.H., dan Kumar V., (1999), "CHAMELEON : A Hierarchical Clustering Algorithm Using Dynamic Modeling", *Proceedings IEEE Transaction on Computer*, vol. 32, hal. 68-75.

[6] Moosinghe H.D.K., dan Pang-Ning T. (2006), "Outlier Detection Using Random Walks", Department of Computer Science & Engineering Michigan State University.

[7] Qing Zang dkk. (2004), "Cluster Cores-based Clustering for High Dimensional Data", Department of Computer Science, Hongkong university of Science & Technology